Journal of Agricultural Machinery

Homepage: https://jame.um.ac.ir

Research Article
Vol. 12, No. 4, 2022, p. 543-558

A Data-driven Model for Predicting the Yield of Recoverable Sugar from Sugarcane

F. Nadernezhad[1], D. M. Imani[2*], M. R. Rasouli[3]

1- MSc Student, Productivity Management Department, School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
2- Assistant Professor, Productivity Management Department, School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
3- Assistant Professor, System and e-Commerce Engineering Department, School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
(*- Corresponding Author Email: Imanim@iust.ac.ir)
https://doi.org/10.22067/jam.2021.69805.1034

Received: 12-04-2021 Revised: 09-06-2021 Accepted: 14-07-2021 Available Online: 14-07-2021	**How to cite this article:** Nadernezhad, F., Imani, D.M., & Rasouli, M.R. (2022). A Data-driven Model for Predicting the Yield of Recoverable Sugar from Sugarcane. *Journal of Agricultural Machinery*, *12*(4), 543-558. (in Persian with English abstract). https://doi.org/10.22067/jam.2021.69805.1034

Introduction

Sugarcane is a strategic agricultural product and increasing productivity and self-sufficiency in its production is of special importance. The most important product of sugarcane is sugar. Various factors like climatic and management conditions affect the yield of sugarcane and recoverable sugar. Crop yield forecasting is one of the most important topics in precision agriculture, which is used to estimate yield, match product supply with demand and manage product to increase productivity. The purpose of this study is to predict and model the factors affecting sugar extracted from sugarcane (recoverable sugar) in the farms of Amir-Kabir sugarcane agro-industry Company of Khuzestan province using machine learning methods.

Materials and Methods

To conduct this study, data from the agro-industrial company Amir-Kabir in the province of Khuzestan from 2010 to 2017 were used. This data has 3223 records which include four sets of data: climate, soil, crop and farm management. This data includes continuous and discrete variables. Discrete variables include production management, soil type, farm, variety, age (cane class), the month of harvest and times irrigation. Continuous variables include area, chemical fertilizer consumption, water consumption per hectare, total water consumption, drain, crop season duration, yield (cane yield) soil EC, purity, time interval drying off to crop harvest, precipitation, min and max temperature, min and max relative humidity, wind speed and evaporation. The recoverable sugar variable is considered as the target variable and is divided into two classes, values greater than or equal to 9 are in the optimal class and less than 9 are in the undesirable class. The other variables are considered as predictor variables. For modeling using the Holdout method the data were randomly divided into two independent sets, a training set and a test set. 70% of the data which includes 2256 records were used for training and 30% of the data which includes 967 records were used for testing. The modeling of this study was performed with the Python programming language version 3.8.6 in the Jupyter notebook environment. Random Forest, Adaboost, XGBoost and SVM (support vector machine) algorithms were used for modeling.

Results and Discussion

To evaluate the models, metrics of accuracy, precision, recall, f1 score and k-fold cross validation were used. The XGBoost model with 94.8% accuracy on the training set and the Adaboost model with 92.4% accuracy on the test set, are the best models. Based on precision and recall metrics Adaboost model with 87% precision and SVM model with 87% recall have better performance than the other models. Based on Repeated 10-fold stratified cross validation using two repeats the SVM model with 92.3% accuracy is the best model. The variables of purity, time interval drying off to crop harvest and crop season duration are the most important variables in predicting the recoverable sugar.

Conclusion

In this study a new approach based on machine learning methods for predicting recoverable sugar from sugarcane was presented. The most important innovation of this study is the simultaneous consideration of management and climatic factors, along with other factors such as soil and crop characteristics for modeling and classification the recoverable sugar percentage from sugarcane. The results show that the performance of all models is acceptable and machine learning methods and ensemble learning algorithms can be used to predict crop yield. The results of this study and the analysis of the rules obtained from the set of decision trees made in the random forest model can be used for managers of different agro-industries in determining appropriate strategies and preparing the conditions to achieve optimal production.

For future research as well as policy making and decision making Amir-Kabir sugarcane agro-industry Company the following suggestions are offered: more samples can be used to obtain more reliable results. Also can be used Deep learning methods, time series analysis and image processing. Use of IOT equipment to collect and real-time processing data on Amir-Kabir sugarcane agro-industry farms.

Keywords: Classification, Machine learning, Modeling, Precision agriculture

نشریه ماشین‌های کشاورزی
https://jame.um.ac.ir

مقاله پژوهشی

جلد ۱۲، شماره ۴، زمستان ۱۴۰۱، ص ۵۵۸-۵۴۳

ارائه مدلی داده‌رانه برای پیش‌بینی عملکرد شکر استحصالی از نیشکر

فاطمه نادرنژاد[۱]، دین محمد ایمانی[۲*]، محمدرضا رسولی[۳]

تاریخ دریافت: ۱۴۰۰/۰۱/۲۳
تاریخ پذیرش: ۱۴۰۰/۰۴/۲۳

چکیده

پیش‌بینی عملکرد محصول یکی از مسائل مهم در حوزه‌ی کشاورزی می‌باشد و به عوامل مختلفی از جمله شرایط آب‌وهوایی، ویژگی‌های خاک، ویژگی‌های محصول و برنامه‌های مدیریتی وابسته می‌باشد. پیش‌بینی دقیق عملکرد محصول می‌تواند در تصمیم‌گیری‌ها و بهینه‌سازی فرآیندها به کشاورزان و صنایع وابسته به کشاورزی کمک نماید و در نهایت منجر به افزایش تولید شود. نیشکر یکی از مهم‌ترین محصولات استراتژیک کشاورزی و منبع تأمین شکر در جهان می‌باشد. هدف پژوهش حاضر پیش‌بینی و بررسی عوامل مؤثر بر میزان شکر استحصالی از نیشکر در مزارع شرکت کشت‌وصنعت نیشکر امیرکبیر با استفاده از الگوریتم‌های یادگیری ماشین می‌باشد. داده‌های جمع‌آوری شده برای این پژوهش مربوط به بازه زمانی سال‌های ۱۳۹۶-۱۳۸۹ شامل ۳۲۲۳ نمونه می‌باشد که شامل چهار مجموعه داده آب‌وهوایی، محصول، خاک و مدیریت مزرعه می‌باشد. برای مدل‌سازی پژوهش از الگوریتم‌های جنگل تصادفی، آدابوست، تقویت گرادیان حداکثری و ماشین بردار پشتیبان استفاده شده و در محیط ژوپیترنوت‌بوک پایتون پیاده‌سازی شده‌اند. مدل جنگل تصادفی با صحت ۹۲/۲٪ برای پیش‌بینی شکر استحصالی در بین مدل‌های ارائه شده بهترین عملکرد را دارد.

واژه‌های کلیدی: طبقه‌بندی، کشاورزی دقیق، مدل‌سازی، یادگیری ماشین

مقدمه

کشاورزی نقش مهمی در اقتصاد جهانی دارد و موادغذایی مورد نیاز انسان را تأمین می‌کند. با افزایش روزافزون جمعیت جهان و به دنبال آن افزایش تقاضا برای موادغذایی فشار بیشتری بر سیستم کشاورزی و منابع طبیعی وارد می‌شود. ورود تکنولوژی‌های جدید در بخش کشاورزی در طی قرن گذشته و در طول انقلاب سبز، به کشاوزی کمک کرده که همگام با تقاضای در حال رشد مواد غذایی و سایر محصولات باشند. تکنولوژی‌ها و رویکردهای جدید می‌توانند اثرات زیست‌محیطی کشاورزی را شناسایی و با حفظ یا کاهش آن، نیازهای غذایی آینده را برطرف کنند. فناوری‌های نوظهوری همانند اینترنت اشیاء، تجزیه و تحلیل کلان داده‌ها و هوش مصنوعی می‌توانند در تصمیم‌گیری‌های آگاهانه مدیریتی با هدف افزایش تولید محصولات استفاده شوند (Sishodia, Ray, & Singh, 2020).

کشاورزی دقیق که امروزه کشاورزی دیجیتال نامیده می‌شود مستلزم استفاده از مجموعه‌ای از این فناوری‌ها برای بهینه‌سازی نهاده‌های کشاورزی برای افزایش میزان تولید کشاورزی و کاهش اتلاف‌ها می‌باشد. کشاورزی دقیق زمینه‌های علمی جدیدی را به‌وجود آورده که با استفاده از رویکردهای داده‌محور منجر به افزایش بهره‌وری در کشاورزی شده و اثرات زیست‌محیطی آن را به حداقل می‌رساند. داده‌های تولید شده در عملیات کشاورزی مدرن توسط انواع مختلفی از حس‌گرها جمع‌آوری می‌شود که درک بهتری از محیط عملیاتی محصول، شرایط خاک، شرایط آب‌وهوایی و داده‌های مربوط به عملیات ماشین‌های کشاورزی را ایجاد می‌کند و منجر به تصمیم‌گیری‌های دقیق‌تر و سریع‌تری می‌شود (Liakos, Busato, Moshou, Pearson, & Bochtis, 2018). پیش‌بینی عملکرد محصول یکی از مهم‌ترین موضوعات در کشاورزی دقیق است، که برای نظارت بر عملکرد، تخمین عملکرد، تطابق عرضه محصول با تقاضا و مدیریت محصول برای افزایش بهره‌وری، استفاده می‌شود و از اهمیت بالایی برخوردار می‌باشد (Liakos *et al.*, 2018). همچنین پیش‌بینی عملکرد محصول یکی از مسائل چالش برانگیز در کشاورزی دقیق می‌باشد و تاکنون مدل‌های زیادی برای آن ارائه و تایید شده است. این مسئله نیاز به استفاده از چندین مجموعه داده دارد، زیرا عملکرد محصول به عوامل مختلفی از جمله شرایط اقلیمی، ویژگی‌های خاک، کودهای شیمیایی و نوع محصول بستگی دارد.

۱- دانشجوی کارشناسی ارشد، گروه مدیریت بهره‌وری، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران
۲- استادیار، گروه مدیریت بهره‌وری، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران
۳- استادیار، گروه مهندسی سیستم‌های هوشمند، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران
(*- نویسنده مسئول: Email: Imanim@iust.ac.ir)
https://doi.org/10.22067/jam.2021.69805.1034

بنابراین می‌توان گفت پیش‌بینی عملکرد محصول کار ساده‌ای نیست و مراحل پیچیده‌ای دارد که باید بر اساس داده‌های موجود حل شود (Van Klompenburg, Kassahun, & Catal, 2020).

ابزارهای داده‌کاوی و یادگیری ماشین می‌توانند داده‌ها را تحلیل و عوامل مؤثر بر عملکرد محصول را شناسایی کنند. داده‌کاوی فرآیند شناسایی و استخراج اطلاعات مهم و مفید و الگوهای پنهان از مجموعه داده‌ها و تبدیل آن‌ها به اطلاعات قابل فهم می‌باشد (Ramesh & Vardhan, 2013). یادگیری ماشین شاخه‌ای از هوش مصنوعی می‌باشد که تمرکز آن ایجاد توانایی یادگیری در ماشین‌ها بدون دخالت انسان و برنامه‌ریزی دقیق می‌باشد. فلسفه یادگیری ماشین بر این است که آینده به گذشته بسیار نزدیک می‌باشد، بنابراین مدل‌ها بر اساس داده‌های گذشته ساخته و آموزش داده می‌شوند و بر اساس آن‌ها آینده پیش‌بینی می‌شود. مدل‌های یادگیری ماشین باتوجه به هدف مسئله می‌توانند پیش‌بینی‌کننده یا توصیفی باشند. برای کسب دانش از داده‌های جمع‌آوری شده و توضیح آن‌چه اتفاق افتاده است از مدل‌های توصیفی استفاده می‌شود، در حالی که از مدل‌های پیش‌بینانه برای پیش‌بینی آینده استفاده می‌شود (van Klompenburg *et al.*, 2020).

نیشکر یکی از محصولات استراتژیک کشاورزی و یکی از مهم‌ترین گیاهان قندی در جهان محسوب می‌شود. این گیاه پتانسیل تولید شکر با کیفیت بالا و به مقدار زیاد در واحد سطح زمین را دارد. اصلی‌ترین محصولی که از نیشکر استحصال می‌شود شکر است. مقدار شکر استحصال شده برای هر رقم نیشکر با توجه به شرایط جغرافیایی و اقلیمی متفاوت می‌باشد. اما معمولاً بین ۱۰ تا ۱۲ درصد برای هر تن نی متفاوت است. قیمت بسیار ارزان شکر در مقایسه با مقدار کالری که ایجاد می‌نماید، شکر را به‌عنوان یکی از منابع غذایی انسان تبدیل کرده و نقش مهمی را در سبد مواد غذایی ضروری مردم جهان دارد. به طوری که ۵/۲ درصد از کل تولیدات غذایی جهان به نیشکر و چغندر قند اختصاص دارد. در مقیاس جهانی عملکرد قند چغندر قند و قند نیشکر در واحد سطح تقریباً مساوی است، اما هزینه تولید نیشکر معمولاً کمتر از چغندر قند می‌باشد (Shooshtari, Ahmadian, & Asfiaa, 2008). بر اساس آمارهای موجود، شکر تولید شده در جهان تقریباً ۸۰ درصد از نیشکر و ۲۰ درصد از چغندر قند به‌دست می‌آید (The Sugar Market, n.d.; Walton, 2020).

با توجه به این‌که نیشکر در مناطق گرمسیری و نیمه‌گرمسیری دنیا و در حوالی مدار ۲۶ درجه تا ۳۴ درجه و ۴۵ دقیقه شمالی می‌روید منطقه جنوب ایران و استان خوزستان منطقه مساعد برای کشت این گیاه محسوب می‌شود. نیشکر در خوزستان در واحدهای هفت‌گانه شرکت توسعه نیشکر و صنایع جانبی و همچنین شرکت‌های کارون، هفت تپه و میان آب کشت می‌شود. این شرکت‌ها و واحدهای کشاورزی وابسته به آن‌ها برای تعیین استراتژی‌های مناسب مدیریتی در زمینه واردات، صادرات و قیمت‌گذاری محصول نیازمند اطلاعاتی پیرامون عملکرد محصول هستند. با استفاده از روش‌های داده‌کاوی و یادگیری ماشین می‌توان داده‌های مربوط به عملکرد محصول و عوامل مؤثر بر آن را شناسایی و بررسی نمود و مدل‌هایی برای پیش‌بینی عملکرد آن ارائه داد که می‌تواند در تصمیم‌گیری‌های بلندمدت استفاده شود و رشد و توسعه‌ی اقتصادی و افزایش تولید را به ارمغان آورد. هدف این پژوهش ارائه‌ی مدلی برای پیش‌بینی شکر استحصالی از نیشکر و بررسی متغیرهای مؤثر بر آن، با استفاده از الگوریتم‌های یادگیری ماشین می‌باشد. مهم‌ترین نوآوری این مطالعه درنظر گرفتن هم‌زمان عوامل مدیریتی و آب‌وهوایی در کنار سایر عوامل از جمله ویژگی‌های خاک و محصول برای مدل‌سازی و پیش‌بینی شکر استحصالی از نیشکر می‌باشد. برای این پژوهش از مجموعه داده‌های شرکت کشت‌وصنعت نیشکر امیرکبیر در بازه زمانی سال‌های ۱۳۹۶-۱۳۸۹ که شامل ۳۲۲۳ نمونه می‌باشد، استفاده شده است.

در پژوهش (Veenadhari, Misra, & Singh, 2011) تأثیر پارامترهای آب‌وهوایی بر میزان بهره‌وری محصول سویا بررسی شده است. در این تحقیق از روش‌های درخت تصمیم (ID3) برای پیش بینی تأثیر پارامترهای آب‌وهوایی استفاده شده است. تحلیل‌های درخت تصمیم نشان می‌دهد که بهره‌وری و عملکرد محصول سویا به‌طور عمده تحت تأثیر رطوبت نسبی، دما و بارندگی می‌باشد.

در پژوهش (Veenadhari, Misra, & Singh, 2014) از رویکردهای یادگیری ماشین برای پیش‌بینی رشد محصول بر اساس پارامترهای آب‌وهوایی استفاده شده است. در این تحقیق نرم‌افزاری با عنوان "Crop Advisor" با کمک الگوریتم C4.5 تأثیر پارامترهای آب‌وهوایی بر عملکرد محصول را بررسی می‌کند و پارامتری که بیش ترین تأثیر را بر عملکرد محصول انتخاب شده دارد مشخص می‌کند.

در پژوهش (Thuankaewsing, Khamjan, Piewthongngam, & Pathumnakul, 2015) راه‌حلی برای مسئله زمان‌بندی برداشت نیشکر، برای گروهی از تأمین‌کنندگان که نیشکر کارخانه شکر در کشور تایلند را تأمین می‌کردند، ارائه شد. برای پیش‌بینی عملکرد نیشکر از شبکه‌های عصبی مصنوعی[1] استفاده شد. برای مدل‌سازی از متغیرهای مختلفی از جمله رقم محصول، نوع خاک و میانگین حداقل و حداکثر دمای روزانه استفاده شده است.

در پژوهش (de Oliveira, Bocca, & Rodrigues, 2017) از سه تکنیک یادگیری ماشین رگرسیون بردار پشتیبان، جنگل تصادفی[2]

1- Artificial neural networks
2- Random forest

و درخت‌های رگرسیون برای پیش‌بینی میزان شکر استحصالی از نیشکر برداشت شده استفاده شده است. نتایج نشان می دهد که مدل جنگل تصادفی با کمترین مقدار خطا بهترین روش برای پیش بینی می‌باشد.

در پژوهش (Balakrishnan *et al.*, 2016) از مجموعه‌ی داده های هواشناسی شامل: میانگین دما، تراکم ابر، دمای روزانه، حداکثر و حداقل دما، تبخیر و تعرق بالقوه، تبخیر و تعرق محصول، فشار بخار و بارندگی برای پیش‌بینی عملکرد محصولات برنج، پنبه، نیشکر، بادام زمینی و ماش سیاه استفاده شده است. الگوریتم های استفاده شده برای پیش‌بینی ماشین بردار پشتیبان[1] و بیز ساده[2] و روش‌های جمعی AdaSVM و AdaNaive می‌باشد. نتایج نشان می‌دهد که روش‌های جمعی AdaSVM و روش AdaNaive نسبت به روش‌های ماشین بردار پشتیبان و بیز قابل قبول‌تر هستند.

در پژوهش (Rajeswari, Suthendran, & Rajakumar, 2017) از دستگاه‌های اینترنت اشیاء برای جمع‌آوری داده‌های کشاورزی و ذخیره‌سازی آن‌ها در فضای ابری استفاده شده است. سپس بر مبنای روش‌های داده‌کاوی پیش‌بینی‌هایی انجام می‌شود. هدف نهایی این تحقیق افزایش تولید محصول و کنترل هزینه‌های تولیدی کشاورزی با استفاده از اطلاعات به‌دست آمده از پیش بینی می‌باشد. مدل هوشمند کشاورزی پیشنهادی در این تحقیق عملکرد محصول را پیش‌بینی می‌کند و تصمیم‌گیری در مورد توالی بهتر محصول بر اساس توالی گذشته محصول در همان مزرعه را با توجه به اطلاعات فعلی مواد مغذی خاک انجام می‌دهد.

در پژوهش (Pande, Purohit, Jadhav, & Shah, 2019) سیستم بهینه پیش‌بینی عملکرد محصول با کمک روش‌های داده‌کاوی پیشنهاد شده است. در این تحقیق برنامه مبتنی بر وب توسعه داده شده که به کشاورزان برای انتخاب مناسب‌ترین محصول برای کشت کمک می کند. این سیستم در مقایسه با سیستم‌های قبلی بهتر است و پارامترهای بیشتری را جهت انتخاب محصول بررسی می‌کند. پارامترهای مورد بررسی شامل رنگ خاک، رطوبت خاک، میزان PH خاک، فصل، بارندگی، دما و آبیاری می‌باشد. بر مبنای این پارامترها و با کمک الگوریتم درخت تصمیم ID3 سیستم مناسب‌ترین محصول را پیشنهاد می‌دهد.

در پژوهش‌های (Ferraro, Rivero, & Ghersa, 2009; Bocca & Rodrigues, 2016; Everingham, Sexton, Skocaj, & Inman-Bamber, 2016; Charoen-Ung & Mittrapiyanuruk, 2018; Medar, Rajpurohit, & Ambekar, 2019; Zakidizaji, Bahrami, Monjezi, & Shiekhdavoodi, 2019) مدل‌هایی بر مبنای روش‌های یادگیری ماشین برای پیش‌بینی عملکرد نیشکر ارائه شده است. جدول ۱ مقاله‌های بررسی شده در این پژوهش را نشان می‌دهد. همان‌طور که مشاهده می‌شود در بیشتر مدل‌ها تمرکز بر پیش‌بینی عملکرد نیشکر می‌باشد و به پیش‌بینی شکر حاصل از آن کمتر پرداخته شده است، همچنین تمامی عوامل مؤثر از جمله عوامل مدیریتی و آب‌وهوایی در کنار هم بررسی نشده‌اند، بنابراین می‌تواند موضوع پژوهش‌های آینده و از جمله شکاف‌های تحقیقاتی محسوب شود. در پژوهش‌های بررسی شده الگوریتم‌های بوستینگ کمتر از سایر الگوریتم‌ها استفاده شده‌اند و با توجه به این‌که از خانواده یادگیری گروهی هستند می‌توانند از الگوریتم‌های تکی عملکرد بهتری داشته باشند و در مدل‌سازی استفاده شوند. در پژوهش حاضر مدلی برای پیش‌بینی عملکرد شکر استحصالی از نیشکر با در نظر گرفتن عوامل مختلفی از جمله عوامل مدیریتی و آب‌وهوایی با استفاده از رویکردهای یادگیری ماشین ارائه شده است.

مواد و روش‌ها

مراحل این پژوهش بر اساس گام‌های متدولوژی کریسپ[3] انجام شده است. به‌طور معمول از متدولوژی کریسپ برای انجام پروژه‌های صنعتی و سازمانی استفاده می‌شود، که به معنی فرآیندهای استاندارد صنعتی متقابل برای داده‌کاوی می‌باشد و در واقع چرخه حیات یک پروژه را نشان می‌دهد. این متدولوژی از شش گام شامل درک موضوع کسب وکار، درک و شناخت داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی و توسعه تشکیل شده است. برای انجام پژوهش از داده‌های مزارع شرکت کشت‌وصنعت نیشکر امیرکبیر استفاده شده است. این مزارع در ۴۵ کیلومتری جنوب اهواز و حدفاصل رودخانه کارون جاده اهواز خرم شهر بین طول های جغرافیایی ′۱۰ و ۴۸ تا ′۲۲ و ۴۸ شرقی و بین عرض‌های ′۵۰ و۳۰ تا ′۵۰ و ۳۱ شمالی واقع هستند و ارتفاع از سطح دریا آن ۷ متر می‌باشد. این شرکت دارای ۴۸۰ مزرعه ۲۵/۵ هکتاری می‌باشد. داده‌های جمع‌آوری شده در بازه زمانی سال‌های ۱۳۹۶-۱۳۸۹ می‌باشد. مجموعه داده‌ها شامل ۳۲۲۳ نمونه (رکورد) می‌باشد که از یکپارچه‌سازی چهار مجموعه داده شامل داده‌های هواشناسی، داده‌های محصول، داده‌های خاک و داده‌های مدیریت مزرعه تشکیل شده است. جدول ۲ توصیف متغیرهای (ویژگی‌های) گسسته و جدول ۳ توصیف متغیرهای (ویژگی‌های) پیوسته را نشان می‌دهد. متغیرهای حداقل و حداکثر دما، حداقل و حداکثر رطوبت نسبی، تبخیر و سرعت باد به‌صورت میانگین در بازه زمانی اردیبهشت-مهر محاسبه شده‌اند، همچنین بارندگی به‌صورت میانگین در بازه زمانی مهر-اردیبهشت محاسبه شده است.

1- Support vector machine
2- Naive Bayes
3- Cross-industry standard process for data mining as CRISP-DM

جدول ۱ – مقاله‌های بررسی شده

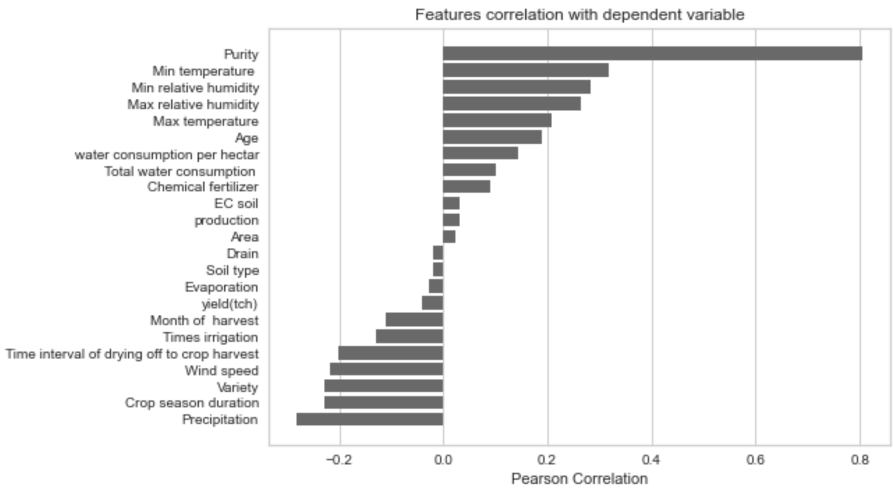
Table 1- Reviewed articles

مرجع **Reference**	محصول **Crop**	تکنیک **Technique**	ابزار **Tool**	ویژگی **Feature**
Ferraro *et al.* (2009)	نیشکر Sugarcane	Classification and regression trees (CART)		مدیریت مزرعه، سن محصول، ماه برداشت، طول فصل زراعی، مساحت، بارندگی Farm management, Month of harvest, Age, Crop season duration, Area, Precipitation
Veenadhari *et al*. 2011	سویا Soybean	Decision tree		رطوبت نسبی، دما، بارندگی، تبخیر Relative Humidity, Temperature, Rainfall, Evaporation
Veenadhari *et al*. (2014)	ذرت، گندم، سویا، برنج Maize, Wheat, Paddy, Soyabean	Decision tree	Web based software in C# language	بارندگی، حداقل و حداکثر دما، تبخیر تعرق بالقوه، پوشش ابر Rainfall, Maximum & Minimum temperature, Potential Evapotranspiration, Cloud cover
Thuankaewsing *et al*. (2015)	نیشکر Sugarcane	Artificial neural networks	Matlab	مزرعه، سن، کلاس محصول، رقم، نوع خاک، آبیاری، طول فصل زراعی، بارندگی، حداقل و حداکثر دما Farm skill, Area, Crop class, Cultivar, Soil type, Irrigation system, Rainfall, Maximum & Minimum temperature, Crop season duration
Balakrishnan *et al*. (2016)	برنج، پنبه، نیشکر و بادام زمینی Rice paddy, Cotton, SugarcaneGroundnut	Support vector machine NaïveBayes-AdaSVM-AdaNaiveBayse	Rapidminer	میانگین و حداقل و حداکثر دما، پوشش ابر، تبخیر تعرق بالقوه، فشار بخار، تعداد روزهای مرطوب، بارندگی Average Temperature, Cloud Cover, Diurnal Temperature, Maximum Temperature, Minimum Temperature, Potential Evapotranspiration, Vapour Pressure, Wet day frequency, Precipitation
Bocca and Rodrigues (2016)	نیشکر Sugarcane	Random forest, Support vector machine, Boosted regression trees, Artificial neural networks, Regression trees		کود شیمیایی، دما، بارندگی، مزرعه، ماه برداشت، رقم محصول، نوع خاک Fertilizer, Temperature, Precipitation, Variety, Age, Soil information, Month of harvest
Everingham *et al*. (2016)	نیشکر Sugarcane	Classification and regression random forest	Statistical software R	بارندگی، دما، تابش خورشید Rainfall, Temperature, Radiation
de Oliveira *et al*. (2017)	شکر استحصالی از نیشکر Recoverable sugar from sugarcane	Support vector regression, Random forests, Regression trees	Statistical software R	اطلاعات خاک، حداقل و حداکثر دما، مجموع تعداد روزها با درجه حرارت منفی و مثبت Soil information, Maximum & Minimum temperature, Sum of degree days, Precipitation, Sum of Negative Degree Days, Fertilization, Variety, Crop season duration
Rajeswari *et al*. (2017)	تصمیم‌گیری بهترین توالی محصول Decide the better crop sequence	Association rule, Decision tree		اطلاعات خاک Soil information
Charoen-Ung, and Mittrapiyanuruk (2018)	نیشکر Sugarcane	Random forest	python	رقم محصول، نوع خاک، مساحت، کود شیمیایی، نوع آبیاری Cane class/type, Soil type, Area, Fertilizer, Rainfall, Epidemic, Water Type
Pande *et al*. (2019)	انتخاب محصول مناسب Selecting suitable crop	Decision tree		رنگ خاک، رطوبت خاک، PH خاک، فصل، بارندگی، آبیاری Soil color, Soil moisture, Soil PH, Season, Rainfall, Irrigation
Medar *et al*. (2019)	نیشکر Sugarcane	Time series, support vector regression	python	دما، دمای نقطه شبنم، دمای خاک، رطوبت خاک، بارندگی، رطوبت نسبی، مدت زمان آفتابی، شاخص NDVI، تبخیر تعرق Temperature, Dew Point Temperature, Soil Temperature, Soil Moisture, Precipitation, Relative Humidity, Sunshine Duration, Evapotranspiration , NDVI
Zakidizaji *et al*. (2019)	نیشکر Sugarcane	C5.0 and QUEST Decision Tree	SPSS	رقم محصول، ماه برداشت، کودشیمیایی، سن گیاه، تعداد دفعات آبیاری، نسبت سطح سمپاشی، بافت خاک، هدایت الکتریکی خاک، مقدار آب مصرفی، زهکشی، مدیریت مزرعه، طول فصل زراعی، مساحت Cultivar, Month of harvest, Chemical fertilizer, Age, Times irrigation, Ratio of surface spraying, Soil texture, Soil electrical conductivity (EC), Water consumption per hectare, Drain, Farm management, Crop duration, Area

جدول ۲- توصیف متغیرهای گسسته پژوهش

Table 2- Description categorical variables used for this study

ردیف **Row**	نام متغیر **Variable name**	توصیف متغیر **Variable description**
1	مدیریت تولید Production management	1, 2
2	بافت خاک Soil type	سندی کلی لوم sandy clay loam، سندی لوم sandy loam سیلتی لوم silty loam ، سیلتی کلیsilty clay سیلتی کلی لوم silty clay loam، سیلتی لومsilty loam کلی لوم clay loam، لوم loam لوم-کلی لوم loam-clay loam
3	مزرعه Farm	480 مزرعه
4	واریته محصول Variety	CP48-103, CP57-614, CP69-1062, CP73-21 IRC0010, IRC0014,IRC9901, IRC9902, IRC9906, MIX, SP70-1143, Tahghighati
5	سن گیاه (کلاس محصول) Age (Cane class)	PC,R1,R2,R3,R4,R5,R6,R7,R8,R9,R10
6	ماه برداشت Month of harvest	فروردین، اردیبهشت، مهر، آبان، آذر، دی، بهمن، اسفند April, may, October, November, December, January, February, March
7	تعداد دفعات آبیاری Times irrigation	11 تا 35 نوبت



شکل ۱- هم‌بستگی بین متغیرهای مستقل با متغیر وابسته (درصد شکر استحصالی)

Fig.1. Features correlation with dependent variable (recoverable sugar)

شکل ۱ هم‌بستگی متغیرهای مستقل (پیش‌بینی‌کننده) با متغیر وابسته یعنی درصد شکر استحصالی را نشان می‌دهد، که بر اساس هم‌بستگی پیرسون[1] محاسبه شده‌اند. همان‌طور که مشاهده می‌شود ویژگی‌های درصد خلوص شربت و حداقل دما همبستگی مثبت بالایی با میزان شکر استحصالی دارند و هم‌چنین متغیر بارندگی بیش‌ترین همبستگی منفی با میزان شکر استحصالی را دارد.

آماده‌سازی داده‌ها

گام سوم در متدولوژی کریسپ آماده‌سازی داده‌ها می‌باشد. در این مرحله متناسب با الگوریتم‌های مورد استفاده برای مدل‌سازی تغییراتی بر روی شکل داده‌ها انجام می‌شود، از جمله تبدیل متغیرهای اسمی و ترتیبی به مقادیر عددی می‌باشد. در این پژوهش متغیرهای بافت خاک، واریته محصول، سن گیاه و ماه برداشت به مقدارهای عددی تبدیل شده‌اند.

1- Pearson

جدول ۳- توصیف متغیرهای پیوسته پژوهش

Table 3- Description continuous variables used for this study

ردیف Number	نام متغیر Variable name	واحد Unit	توصیف متغیر Variable description	میانگین Average
1	مساحت Area	ha	متغیر ورودی Input variable	22.41
2	کود مصرفی Chemical fertilizer consumption	kg ha^{-1}	متغیر ورودی Input variable میزان کل کود مصرفی برای مزارع که طی ۴ مرحله انجام می‌شود.	354
3	مصرف آب در هکتار Water consumption per hectare	m^3 ha^{-1}	متغیر ورودی Input variable	1404
4	کل آب مصرفی Total water consumption	m^{3}ha^{-1}	متغیر ورودی Input variable مساحت قابل برداشت * مصرف آب در هکتار	31902
5	کل زه خروجی Drain	m^{3}ha^{-1}	متغیر ورودی Input variable	15320
6	طول فصل زراعی Crop season duration	day	متغیر ورودی Input variable	391
7	عملکرد Yield	ton ha^{-1}	متغیر ورودی Input variable مقدار نیشکر برداشت شده از مساحت قابل برداشت	71.7
8	هدایت الکتریکی خاک Soil EC	ds m^{-1}	متغیر ورودی Input variable	4.96
9	درصد خلوص شربت Purity	%	متغیر ورودی Input variable	87.20
10	شکر تصفیه شده (شکر استحصالی) Recoverable sugar	%	متغیر هدف Target variable	10.22
11	فاصله زمانی قطع آب تا برداشت محصول Time interval of drying off to crop harvest	day	متغیر ورودی Input variable	88
12	حداقل دما Min temperature	$^{\circ}$C	متغیر ورودی Input variable	23.78
13	حداکثر دما Max temperature	$^{\circ}$C	متغیر ورودی Input variable	42.65
14	بارندگی Mean precipitation	mm	متغیر ورودی Input variable	
15	حداقل رطوبت نسبی Min relative humidity	%	متغیر ورودی Input variable	18.50
16	حداکثر رطوبت نسبی Max relative humidity	%	متغیر ورودی Input variable	57.64
17	سرعت باد Wind speed	m s^{-1}	متغیر ورودی Input variable	5
18	تبخیر Evaporation	mm	متغیر ورودی Input variable	12.98

تبدیل ویژگی‌های عددی به بازه‌های مناسب گسسته‌سازی[1] نامیده می‌شود. در این پژوهش متغیر شکر استحصالی با توجه به نظر خبره شرکت مورد مطالعه به ۲ کلاس مطلوب (۰) و نامطلوب (۱) تقسیم شده است به این ترتیب که مقادیر بیشتر مساوی ۹ برای درصد شکر استحصالی در کلاس مطلوب و مقادیر پایین‌تر از ۹ در کلاس نامطلوب قرار دارند.

با توجه به این‌که متغیرهای مورد استفاده برای مدل‌سازی مقیاس یکسانی ندارند، برای جلوگیری از این‌که مقیاس‌های متفاوت لطمه‌ای

1- Discretization

به فرآیند تحلیل وارد نکند باید نرمال‌سازی شوند. در نرمال‌سازی داده‌ها سعی می‌شود تا وزن یکسانی به کلیه ویژگی‌ها داده شود. روش‌های متعددی برای نرمال‌سازی داده‌ها وجود دارد. در این پژوهش از روش نرمال‌سازی مین-ماکس[۱] استفاده شده است. این روش نرمال‌سازی یک تبدیل خطی را بر روی داده‌های اولیه اجرا می‌کند. در این روش یک مقدار از ویژگی A مانند x_i به مقداری مانند x در محدوده‌ی $[New_{min}, New_{max}]$ با رابطه (۱) نگاشت می‌شود (Han, Kamber, & Pei, 2019):

$$x = \frac{x_i - x_{min}}{x_{max} - x_{min}} * (New_{max} - New_{min}) + New_{min} \quad (۱)$$

که x_{max} و x_{min} به‌ترتیب بیش‌ترین و کم‌ترین مقدار موجود برای ویژگی مورد نظر و New_{max} و New_{min} محدوده جدید است که در اینجا بازه [0,1] می‌باشد.

الگوریتم‌های یادگیری ماشین

برای مدل‌سازی این پژوهش از چهار الگوریتم یادگیری ماشین استفاده شده است که در ادامه معرفی می‌شوند. متغیرهای پیوسته و گسسته که در جدول‌های ۲ و ۳ معرفی شدند برای ساخت مدل‌ها و پیش‌بینی کلاس‌های تعیین شده برای شکر استحصالی از نیشکر استفاده شده است.

ماشین بردار پشتیبان

ماشین بردار پشتیبان به‌طور ذاتی یک دسته‌بند دودویی است که یک خط یا ابر صفحه جداکننده برای طبقه‌بندی داده‌های نمونه می‌سازد. قابلیت طبقه‌بندی ماشین بردار پشتیبان سنتی را می‌توان به‌وسیله‌ی تبدیل فضای ویژگی‌های اصلی به فضای ویژگی با ابعاد بالاتر با استفاده از حقه کرنل[۲] به‌طور اساسی افزایش داد. ماشین بردار پشتیبان برای یادگیری از توابع خطی استفاده می‌کند در برخی از موارد غیرخطی، ماشین بردار پشتیبان از تکنیک کرنل برای نگاشت کردن داده‌ها در فضای ویژگی با ابعاد بالا استفاده می‌کند که توابع خطی می‌توانند به‌کار برده شوند (Liakos *et al.*, 2018; Palanivel & Surianarayanan, 2019).

جنگل تصادفی[۳]

جنگل‌های تصادفی از الگوریتم‌های یادگیری گروهی هستند که می‌توانند برای مسائل طبقه‌بندی و رگرسیون استفاده شوند. مزیت کلیدی جنگل تصادفی این است که می‌تواند روابط غیرخطی و سلسله مراتبی را بین متغیرهای پیش‌بینی‌کننده و متغیر پاسخ با استفاده از رویکرد یادگیری گروهی بررسی کند. مدل جنگل تصادفی گروهی از درختان تصمیم می‌باشد که هر درخت تصمیم با مجموعه‌ای از داده‌ها که به روش نمونه‌برداری با جایگذاری (تکنیک بوت‌استرپ)[۴] انتخاب شده‌اند آموزش داده می‌شود. برای ساخت هر درخت در هر مرحله زیرمجموعه‌ای از ویژگی‌ها به‌صورت تصادفی انتخاب شده سپس بر اساس معیارهای متفاوتی مثل سنجه جینی[۵] بهترین ویژگی برای تفکیک داده‌ها از زیر مجموعه تصادفی از ویژگی‌ها انتخاب می‌شود. این سنجه ناخالصی نمونه‌ها را در مجموعه D بر اساس رابطه (۲) محاسبه می‌کند:

$$\text{Gini(D)} = 1 - \sum_{i=1}^{m} p_i^2 \quad (۲)$$

که در آن p_i احتمال تعلق یک نمونه در مجموعه D را به کلاس C_i نشان می‌دهد و برای هر ویژگی هرچه مقدار آن کمتر باشد یعنی آن ویژگی اطلاعات بیشتری را به ما می‌دهد و برای تفکیک انتخاب می‌شود (Han *et al.*, 2019). برای پیش‌بینی نمونه‌های داده آزمون، ابتدا پیش‌بینی هر درخت انجام می‌شود و سپس همه پیش‌بینی‌ها ادغام شده و رای‌گیری با رویکرد سهل‌گیرانه یا سخت‌گیرانه انجام می‌شود و پیش‌بینی نهایی اعلام می‌شود (Everingham *et al.*, 2016).

بوستینگ[۶]

بوستینگ یکی از تکنیک‌های یادگیری گروهی است که تلاش می‌کند دسته‌بندی قوی از تعدادی دسته‌بند ضعیف (یادگیرنده ضعیف[۷]) ایجاد کند. این کار با ساخت یک مدل از داده‌های آموزش، سپس ایجاد یک مدل دوم که سعی می‌کند خطاهای مدل اول را اصلاح کند، انجام می‌شود. در واقع در بوستینگ برخلاف روش‌های بگینگ[۸] مانند جنگل تصادفی، مدل‌ها به‌طور مستقل ساخته نمی‌شوند بلکه به‌طور متوالی ساخته می‌شوند. آدابوست[۹] اولین الگوریتم موفق بوستینگ است که برای طبقه‌بندی‌های باینری (دودویی) توسعه داده شده است. آدابوست مخفف بوستینگ تطبیقی بوده که توسط یاو فروند و رابرت شاپیر[۱۰] ابداع شد. در واقع آدابوست یک متا الگوریتم[۱۱] است که برای ارتقای عملکرد همراه دیگر الگوریتم‌های یادگیری استفاده می‌شود. در این الگوریتم دسته‌بندی در هر مرحله جدید بر مبنای نمونه‌های غلط طبقه‌بندی شده در مراحل قبل، تنظیم می‌گردد. در الگوریتم آدابوست در هر دور یک دسته‌بند ضعیف اضافه می‌شود.

1- Min-Max
2- Kernel trick
3- Random forest
4- Bootstrap
5- Gini index
6- Boosting
7- Weak learner
8- Bagging
9- Adaptive boosting
10- Freund & Schapire
11- Meta-algorithm

در هر مرحله دسته‌بند پایه فقط کافیست از دسته‌بند تصادفی (۵۰٪) بهتر باشد (Freund and Schapire, 1997).

گرادیان بوستینگ

گرادیان تقویتی[1] یکی از الگوریتم‌های یادگیری ماشین است که برای مسائل طبقه‌بندی و رگرسیون به‌کار می‌رود. یک مدل پیش‌بینی در قالب گروهی از یادگیرنده‌های ضعیف ایجاد می‌کند که معمولاً یادگیرنده‌ها درختان تصمیم هستند. مدل‌سازی همانند دیگر روش‌های بوستینگ به شکل مرحله‌ای می‌باشد. این الگوریتم مشابه بوستینگ تطبیقی می‌باشد اما از جنبه‌های خاصی با آن متفاوت است. در واقع در این روش مسئله بوستینگ (تقویتی) به‌عنوان یک مسئله بهینه‌سازی مطرح می‌شود. یعنی در هر مرحله تابع ضرری در نظر گرفته می‌شود و هدف بهینه‌سازی آن می‌باشد. این ایده اولین بار توسط بریمن[2] توسعه داده شد. الگوریتم تقویت گرادیان حداکثری[3] (XGBoost) از دسته الگوریتم‌های گرادیان تقویتی و الگوریتم‌های گروهی می‌باشد که می‌تواند برای مسائل رگرسیون و طبقه‌بندی استفاده شود. الگوریتم XGBoost به دلیل سرعت بسیار بالا در مقایسه با سایر الگوریتم‌های گرادیان بوستینگ و عملکرد بسیار خوبی که دارد بسیار محبوب است و هم‌چنین در مسابقات یادگیری ماشین استفاده می‌شود (Brownlee, 2020).

مدل‌سازی

مدل‌سازی این تحقیق با زبان برنامه‌نویسی پایتون نسخه ۳،۸،۶ و در محیط ژوپیتر نوت‌بوک انجام شده است. برای مدل‌سازی داده‌ها با تکنیک هلداوت[4] به‌صورت تصادفی به دو مجموعه مستقل آموزشی و آزمایشی تقسیم شده‌اند. معمولاً دو سوم داده‌ها به مجموعه‌ی آموزشی و یک سوم باقی‌مانده به مجموعه‌ی آزمایشی تخصیص داده می‌شود. مجموعه داده‌های آموزشی برای ساخت مدل و مجموعه داده‌های آزمایشی برای ارزشیابی مدل استفاده می‌شود (Han *et al*., 2019). در این پژوهش ۷۰ درصد از داده‌ها برای آموزش و ۳۰ درصد برای آزمایش استفاده شده‌اند. برای تنظیم و به‌دست آوردن مقادیر بهینه هایپرپارامترهای الگوریتم‌ها از روش جست و جوی شبکه‌ای[5] به همراه اعتبارسنجی متقابل استفاده شده است. جدول ۴ هایپرپارامترهای تنظیم شده به همراه مقدارهایشان را نشان می‌دهد.

نتایج و بحث

برای ارزیابی مدل‌های ساخته شده از سنجه‌های صحت، دقت، فراخوانی و امتیاز F1 و هم‌چنین اعتبارسنجی متقابل داده‌ها استفاده شده است. روابط (۳) تا (۶) سنجه‌های استفاده شده را نشان می‌دهند.

$$\text{صحت} = \frac{TP+TN}{P+N} \qquad (۳)$$

$$\text{دقت} = \frac{TP}{TP+FP} \qquad (۴)$$

$$\text{فراخوانی، حساسیت} = \frac{TP}{P} \qquad (۵)$$

$$\text{F1 امتیاز} = \frac{2*Precision*Recall}{Precision+Recall} \qquad (۶)$$

در روابط بالا P و N به‌ترتیب تعداد کل نمونه‌ها در کلاس مثبت و منفی اشاره می‌کند. درست مثبت (TP)[6]، این اصطلاح اشاره به تعداد نمونه‌هایی دارد که در کلاس مثبت هستند و توسط دسته‌بند به‌درستی برچسب خورده است. درست منفی (TN)[7]، این اصطلاح اشاره به تعداد نمونه‌هایی دارد که در کلاس منفی هستند و توسط دسته‌بند به درستی برچسب منفی خورده است. نادرست مثبت (FP)[8]، تعداد نمونه‌هایی هستند که در کلاس منفی قرار دارند اما به نادرست برچسب مثبت خورده‌اند. نادرست منفی (FN)[9]، تعداد نمونه‌هایی هستند که در کلاس مثبت قرار دارند اما به نادرست برچسب منفی خورده‌اند. معیار صحت کارایی کلی مدل را می‌سنجد. سنجه‌های دقت و حساسیت به‌صورت گسترده‌ای در دسته‌بندی استفاده می‌شوند. سنجه دقت به‌عنوان یک سنجه درستی در نظر گرفته می‌شود (درصدی از نمونه‌ها که مثبت برچسب‌گذاری می‌شوند و واقعا کلاس آن‌ها مثبت است) و سنجه فراخوانی یک سنجه تمامیت است (درصدی از نمونه‌های کلاس مثبت که به درستی دسته‌بندی می‌شوند) و با سنجه حساسیت برابر است. سنجه امتیاز F1 میانگین هارمونیک[10] دو سنجه دقت و فراخوانی است. در فرمول آن وزن یکسانی به دقت و فراخوانی تخصیص داده شده است (Han *et al*., 2019). جدول ۵ نتایج مدل‌های ساخته شده برای پیش‌بینی شکر استحصالی را نشان می‌دهد. سنجه‌های دقت، حساسیت و امتیاز F1 برای مجموعه آزمایشی و سنجه صحت برای مجموعه آموزشی و آزمایشی محاسبه شده است. مقادیر جدول بر اساس درصد می‌باشند.

1- Gradient boosting
2- Breiman
3- Extreme gradient boosting
4- Holdout
5- Grid search

6- True positive
7- True negative
8- False positive
9- False negative
10- Harmonic

جدول ٤- هایپرپارامترهای تنظیم‌شده الگوریتم ها
Table 4- Tuned of algorithms hyperparameters

الگوریتم Algorithms	مقدار تنظیم‌شده هایپرپارامترها با روش جست‌وجوی شبکه‌ای به‌همراه اعتبارسنجی متقابل Hyperparameters by cross-validated grid-search
جنگل تصادفی Random forest	تعداد درختان : تعداد درختان تصمیم در جنگل ۱۰۰ The number of trees in the forest. The maximum depth of the tree حداکثر عمق درخت: ۷ حداقل نمونه در گره برگ: کم‌ترین نمونه لازم در گره که برگ محسوب شود. ۱۲ The minimum number of samples required to be at a leaf node حداکثر ویژگی‌ها: بیش‌ترین تعداد ویژگی‌هایی که هنگام جست‌وجوی بهترین تقسیم باید در نظر گرفته شود، ۸ The number of features to consider when looking for the best split
آدابوست AdaBoost	تعداد مدل‌ها: تعداد مدل‌هایی که فرآیند آموزش را تکرار می‌کنند. ۵۰ The maximum number of estimators at which boosting is terminated نرخ یادگیری: سهم هر طبقه‌بند در وزن‌ها در هر دور را کنترل می‌کند. توازنی بین میزان یادگیری و تعداد دورها (طبقه‌بندها) وجود دارد. ۰,۲۵ Learning rate shrinks the contribution of each classifier by learning_rate
تقویت گرادیان حداکثری XGBoost	تعداد مدل‌ها: تعداد درختان تصمیم در جنگل (تعداد دفعات یادگیری مدل) ۴۵ Number of boosting rounds حداکثر عمق درخت: ۳ نرخ یادگیری: کنترل وزن مدل‌ها در هر دور، ۰,۱۷ Learning rate
ماشین بردار پشتیبان SVC	کرنل: هسته، جداسازی کلاس‌ها در فضای داده‌ها، خطی Specifies the kernel type to be used in the algorithm هایپرپارامتر جریمه C: پارامتر تنظیم میزان جریمه‌ای که به داده‌هایی که اشتباه دسته‌بندی شده‌اند داده می‌شود. ۱۵ Regularization parameter

جدول ۵- نتایج ارزیابی مدل‌ها
Table 5- Results of models evaluation

مدل Model	صحت مجموعه آموزش Accuracy train	صحت مجموعه آزمایش Accuracy test	دقت Precision	فراخوانی Recall	امتیاز F1 F1 score
جنگل تصادفی RF	93.7	92.3	85	68	75
آدابوست AdaBoost	92.6	92.4	87	67	76
تقویت گرادیان حداکثری XGBoost	94.8	92.1	81	72	76
ماشین بردار پشتیبان SVC	92.5	92	65	87	74

همان‌طور که جدول ۵ نشان می‌دهد بـر اسـاس سـنجه صـحت امتیاز مدل‌ها بسیار نزدیک به‌هم و با تفاوت نـاچیزی مـدل آدابوسـت عملکرد بهتری دارد. هم‌چنین بر اساس معیارهـای دقـت و فراخـوانی به‌ترتیب مدل‌های آدابوست و ماشین بردار پشـتیبان عملکـرد بهتـری نسبت به سایر مدل‌ها دارند. بر اساس معیار F1 مدل‌های آدابوسـت و تقویت گرادیان حداکثری با تفاوت ناچیزی نسبت به مدل‌هـای دیگـر عملکرد بهتری دارند.

اعتبارسنجی متقابل

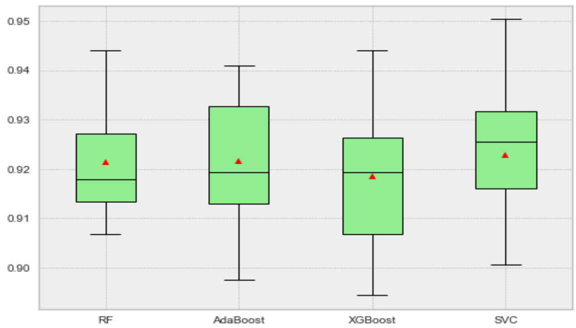
اعتبارسنجی متقابل یکی از روش‌های ارزیابی مدل می‌باشـد کـه تعیین می‌کند نتایج مدل ساخته شده بر روی یک مجموعه داده تا چه اندازه قابل تعمیم و مستقل از داده‌های آموزشی است. در اعتبارسنجی متقابل[1] که همراه با پارمتر k استفاده می‌شود، داده‌های اولیه به‌صورت تصادفی به k زیرمجموعه تقسیم می‌شوند. هر دفعه یک زیر مجموعه برای اعتبارسنجی و k-1 تای دیگر برای آمـوزش بـه‌کـار مـی‌رود. در نهایت میانگین ایـن k اعتبارسـنجی بـه‌عنـوان امتیـاز نهـایی اعـلام می‌شود.

1- K-fold cross-validation

جدول ٦– نتایج اعتبارسنجی متقابل

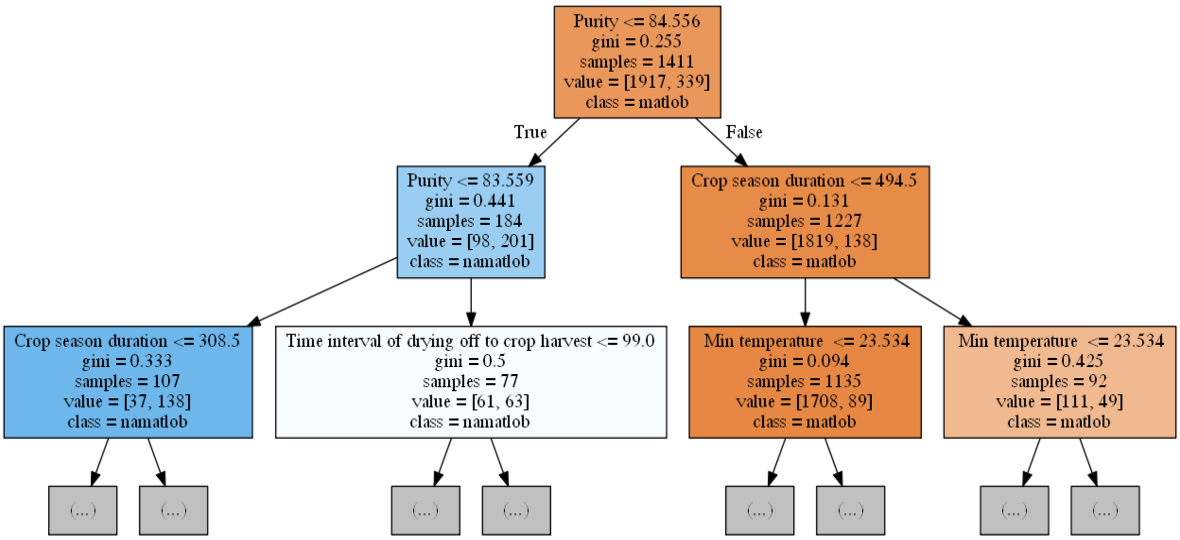
Table 6- Result of 10-fold cross-validation

مدل Model	صحت Accuracy	انحراف استاندارد **Standard deviation**
جنگل تصادفی RF	92.1	0.011
آدابوست AdaBoost	92.2	0.012
تقویت گرادیان حداکثری XGBoost	91.9	0.013
ماشین بردار پشتیبان SVC	92.3	0.013



شکل ۲– نمودار جعبه‌ای اعتبارسنجی متقابل مدل‌ها

Fig.2. Boxplot of CV scores of all classifiers over 10-fold stratified cross validation



شکل ۳– بخشی از درخت تصمیم شماره ۱ از مجموعه درختان تصمیم جنگل تصادفی

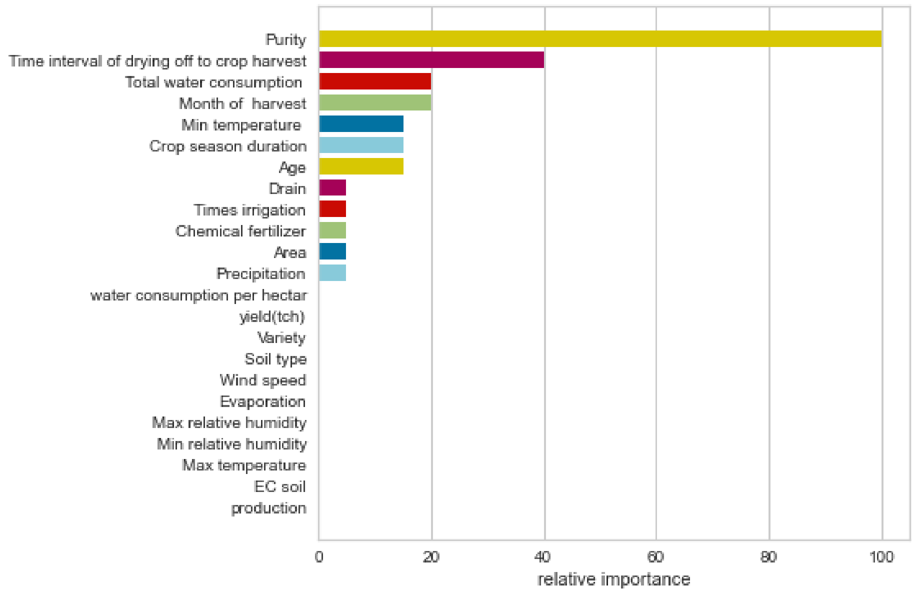
Fig.3. Part of decision tree No.1 from the Random Forest decision trees collection

جدول ۷– قوانین استخراجی از درخت تصمیم شماره ۱ از مجموعه درختان تصمیم جنگل تصادفی

Table 7- Extraction rules from decision tree No.1, from the Random Forest decision trees collection

```
|-- Purity >  84.556
|   |-- Crop season duration <= 494.500
|   |   |-- Min temperature  <= 23.534
|   |   |   |-- Crop season duration <= 376.500
|   |   |   |   |-- truncated branch of depth 4
|   |   |   |-- Crop season duration >  376.500
|   |   |   |   |-- truncated branch of depth 4
|   |   |-- Min temperature  >  23.534
|   |   |   |-- Precipitation <= 0.376
|   |   |   |   |-- class: 0.0
|   |   |   |-- Precipitation >  0.376
|   |   |   |   |-- truncated branch of depth 4
|   |-- Crop season duration >  494.500
|   |   |-- Min temperature  <= 23.534
|   |   |   |-- Total water consumption  <= 30014.873
|   |   |   |   |-- class: 0.0
|   |   |   |-- Total water consumption  >  30014.873
|   |   |   |   |-- truncated branch of depth 3
|   |   |-- Min temperature  >  23.534
|   |   |   |-- Times irrigation <= 28.500
|   |   |   |   |-- truncated branch of depth 2
|   |   |   |-- Times irrigation >  28.500
|   |   |   |   |-- class: 0.0
```

```
|-- Purity <= 84.556
|   |-- Purity <= 83.559
|   |   |-- Crop season duration <= 308.500
|   |   |   |-- class: 0.0
|   |   |-- Crop season duration >  308.500
|   |   |   |-- water consumption per hectar <= 1397.852
|   |   |   |   |-- truncated branch of depth 3
|   |   |   |-- water consumption per hectar >  1397.852
|   |   |   |   |-- truncated branch of depth 2
|   |-- Purity >  83.559
|   |   |-- Time interval of drying off to crop harvest <
= 99.000
|   |   |   |-- yield(tch)  <= 74.989
|   |   |   |   |-- truncated branch of depth 2
|   |   |   |-- yield(tch)  >  74.989
|   |   |   |   |-- truncated branch of depth 2
|   |   |-- Time interval of drying off to crop harvest >
99.000
|   |   |   |-- Max temperature <= 42.952
|   |   |   |   |-- class: 1.0
|   |   |   |-- Max temperature >  42.952
|   |   |   |   |-- class: 1.0
```



شکل ٤– اهمیت متغیرها در مدل آدابوست

Fig.4. Feature importance of 23 features using AdaBoost Classifier

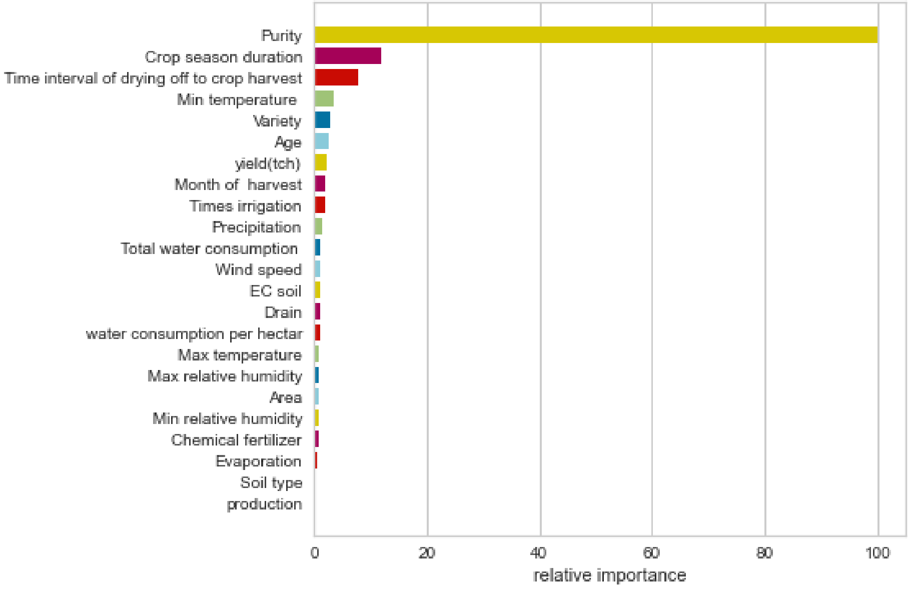
اگر زیرمجموعه‌ها یا بخش‌ها به گونه‌ای انتخاب شوند که توزیع کلاس نمونه‌ها در هر بخش به صورت تقریبی یکسان باشد، اعتبارسنجی متقابل و طبقه‌بندی شده[1] نام دارد (Han *et al.*, 2019). در این پژوهش برای ارزیابی مدل‌های ارائه شده برای پیش‌بینی شکر استحصالی، از اعتبارسنجی متقابل و طبقه‌بندی شده داده‌ها با ۱۰ لایه و ۲ تکرار و معیار صحت برای ارزیابی مدل‌ها استفاده شده است. جدول ۶ نتایج اعتبارسنجی متقابل مدل‌ها را برای پیش‌بینی شکر استحصالی بر اساس درصد نشان می‌دهد. بر این اساس مدل ماشین بردار پشتیبان با صحت ۹۲/۳٪ بهترین مدل برای پیش‌بینی شکر استحصالی می‌باشد. شکل ۲ نمودار جعبه‌ای ارزیابی مدل‌ها با روش اعتبارسنجی متقابل را نشان می‌دهد.

شکل ۳ سه سطح از درخت تصمیم شماره یک از مجموعه ۱۰۰ درخت ساخته شده در مدل جنگل تصادفی را نشان می‌دهد. از شاخص جینی برای تقسیم‌بندی استفاده شده و ویژگی‌هایی که به ریشه درخت نزدیک‌تر هستند اهمیت بیشتری دارند. هم‌چنین در

1- Stratified Cross-validation

جدول ۷ مجموعه قوانین استخراج شده از این درخت را در سه سطح نمایش می‌دهد.

شکل‌های ۴ و ۵ میزان اهمیت متغیرهای استفاده شده برای مدل‌سازی درصد شکر استحصالی را به‌ترتیب برای مدل‌های آدابوست و جنگل تصادفی نشان می‌دهد. همان‌طور که مشاهده می‌شود متغیرهای درصد خلوص شربت، فاصله زمانی قطع آب تا برداشت محصول، طول فصل زراعی و کل آب مصرفی مهم‌تر از سایر متغیرها می‌باشند.



شکل ۵– اهمیت متغیرها در مدل جنگل تصادفی

Fig.5. Feature importance of 23 features using Random Forest Classifier

نتیجه‌گیری

در این مطالعه رویکرد جدیدی مبتنی بر روش‌های یادگیری ماشین برای پیش‌بینی شکر استحصالی از نیشکر ارائه شد. برای مدل‌سازی داده‌ها به دو دسته آموزشی و آزمایشی تقسیم شدند. ۲۲۵۶ نمونه که ۷۰٪ از داده‌ها را شامل می‌شود در مجموعه آموزشی و ۹۶۷ نمونه که ۳۰٪ داده‌ها می‌باشد برای مجموعه آزمایشی در نظر گرفته شد و سپس مدل‌های ساخته شده با سنجه‌های متفاوتی بر روی مجموعه داده آزمایشی ارزیابی شدند. نتایج نشان می‌دهد که عملکرد همه مدل‌ها قابل قبول می‌باشد و می‌توان از روش‌های یادگیری ماشین و الگوریتم‌های یادگیری جمعی برای پیش‌بینی استفاده نمود. نتایج حاصل از بررسی اهمیت ویژگی‌ها نشان می‌دهد متغیرهای درصد خلوص شربت، طول فصل زراعی، فاصله زمانی قطع آب تا برداشت محصول، ماه برداشت و حداقل دما از متغیرهای مهم و تأثیرگذار بر عملکرد شکر استحصالی می‌باشند. در پژوهش‌های گذشته نتایج حاصل از بررسی متغیرهای مؤثر بر عملکرد شکر استحصالی از نیشکر متفاوت می‌باشند. برای مثال در پژوهش (Ferraro *et al.*, 2009) متغیرهای رقم محصول و سن محصول بیش‌ترین اهمیت را بر عملکرد شکر استحصالی از نیشکر داشته و میزان بارندگی کم‌اهمیت‌ترین متغیر بوده است. در پژوهش دیگری (de Oliveira *et al.*, 2017) متغیرهای آب‌وهوایی از جمله مجموع درجه حرارت روزانه و حداقل دما بیش‌ترین اهمیت را بر عملکرد شکر استحصالی از نیشکر داشته و متغیرهای مربوط به خاک از جمله نوع خاک اهمیت کمتری داشته‌اند. نتایج حاصل از این مطالعه و تحلیل قوانین به‌دست آمده از مجموعه ۱۰۰ درخت تصمیم ساخته شده در مدل جنگل تصادفی می‌تواند برای مدیران کشت‌وصنعت‌های مختلف در تعیین استراتژی های مناسب و آماده سازی شرایط برای دستیابی به تولید مطلوب و بهینه استفاده شود.

پیشنهادها

برای پژوهش‌های آینده و همچنین سیاست‌گذاری و تصمیم‌گیری شرکت کشت‌وصنعت نیشکر امیرکبیر پیشنهادهایی به شرح زیر ارائه می‌شود:

- ✓ در تحقیقات آینده می‌توان از تعداد نمونه‌های بیشتر استفاده شود که نتایج با اطمینان بیشتری حاصل شود.
- ✓ همچنین می‌توان از روش‌های یادگیری عمیق و تحلیل سری‌های زمانی و پردازش تصویر استفاده شود.

- ✓ ترکیب الگوریتم‌های یادگیری ماشین با یک‌دیگر و مقایسه با مدل‌های ارائه شده در پژوهش حاضر به‌منظور دستیابی به مدل‌هایی با صحت بالاتر.
- ✓ استفاده از تجهیزات اینترنت اشیاء برای جمع‌آوری و پردازش بلادرنگ داده‌ها در مزارع کشت‌وصنعت نیشکر امیرکبیر

سپاسگزاری

نویسندگان این مقاله از شرکت کشت و صنعت نیشکر امیرکبیر که داده‌های مورد نیاز برای انجام پژوهش را در اختیار قرار داده‌اند، کمال تشکر را دارند.

References

1. Balakrishnan, N., & Muthukumarasamy, G. (2016). Crop production-ensemble machine learning model for prediction. I*nternational Journal of Computer Science and Software Engineering*, *5*, 148.
2. Bocca, F. F., & Rodrigues, L. H. A. (2016). The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Computers and Electronics in Agriculture*, *128*, 67-76. https://doi.org/10.1016/j.compag.2016.08.015.
3. Brownlee, J. (2020a). How to Develop Your First XGBoost Model in Python with scikit-learn. https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/.
4. Brownlee, J. (2020b). A Gentle Introduction to XGBoost for Applied Machine Learning. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.
5. Brownlee, J. (2020c). Extreme Gradient Boosting (XGBoost) Ensemble in Python. https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/.
6. Charoen-Ung, P., & Mittrapiyanuruk, P. (2018). *Sugarcane Yield Grade Prediction using random forest with forward feature selection and hyper-parameter tuning*. Pages 33-42. International Conference on Computing and Information Technology: Springer. https://doi.org/10.1007/978-3-319-93692-5_4
7. de Oliveira, M. P. G., Bocca, F. F., & Rodrigues, L. H. A. (2017). From spreadsheets to sugar content modeling: A data mining approach. *Computers and Electronics in Agriculture*, *132*, 14-20. https://doi.org/10.1016/j.compag.2016.11.012
8. Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, *36*, 27. https://doi.org/10.1007/s13593-016-0364-z
9. Ferraro, D. O., Rivero, D. E., & Ghersa, C. M. (2009). An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. *Field Crops Research*, *112*, 149-157. https://doi.org/10.1016/j.fcr.2009.02.014
10. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119-139. https://doi.org/10.1006/jcss.1997.1504
11. Han, J., Kamber, M., & Pei, J. (2019). *Data mining: concepts and techniques*, 3rd ed. Niaze danesh. Tehran.
12. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*, 2674. https://doi.org/10.3390/s18082674
13. Medar, R. A., Rajpurohit, V. S., & Ambekar, A. M. (2019). Sugarcane Crop Yield Forecasting Model Using Supervised Machine Learning. *International Journal of Intelligent Systems and Applications*, *11*, 11. https://doi.org/10.5815/ijisa.2019.08.02
14. Palanivel, K., & Surianarayanan, C. (2019). An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology*, *10*, 110-118. https://ssrn.com/abstract=3555087
15. Pande, A., Purohit, S., Jadhav, S., & Shah, K. (2019). Optimum Crop Prediction using Data Mining and Machine Learning Techniques. *International Journal for Research in Applied Science and Engineering Technology*, *7*, 2392-2396. https://doi.org/10.22214/ijraset.2019.3436
16. Rajeswari, S., Suthendran, K., & Rajakumar, K. (2017). *A smart agricultural model by integrating IoT, mobile and cloud-based big data analytics*. Pages 1-5. 2017 International Conference on Intelligent Computing and Control (I2C2): IEEE. https://doi.org/10.1109/I2C2.2017.8321902
17. Ramesh, D., & Vardhan, B. V. (2013). Data mining techniques and applications to agricultural yield data. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*, 3477-3480.
18. Shooshtari, M. B., Ahmadian, S., & Asfiaa, G. (2008). *Sugarcane in Iran*. Aeeizh. Tehran.
19. Sishodia, R. P., Ray, R. L., & Singh, S. K. (2020). Applications of remote sensing in precision agriculture: A review. *Remote Sensing*, *12*, 3136. https://doi.org/10.3390/rs12193136
20. The Sugar Market. (n.d.). About Sugar. Retrieved from https://www.isosugar.org/sugarsector/sugar
21. Thuankaewsing, S., Khamjan, S., Piewthongngam, K., & Pathumnakul, S. (2015). Harvest scheduling algorithm to equalize supplier benefits: A case study from the Thai sugar cane industry. *Computers and Electronics in Agriculture*, *110*, 42-55. https://doi.org/10.1016/j.compag.2014.10.005

22. Van Klompenburg, T., Kassahun, A., & Catal, C. 2020. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture, 177*, 105709. https://doi.org/10.1016/j.compag.2020.105709
23. Veenadhari, S., Misra, B., & Singh, C. (2011). Soybean productivity modelling using decision tree algorithms. *International Journal of Computer Applications*, *27*, 11-15.
24. Veenadhari, S., Misra, B., & Singh, C. (2014). *Machine learning approach for forecasting crop yield based on climatic parameters*. Pages 1-5. 2014 International Conference on Computer Communication and Informatics: IEEE. https://doi.org/10.1109/ICCCI.2014.6921718
25. Walton, J. The 5 Countries That Produce the Most Sugar. https://www.investopedia.com/articles/investing/101615/5-countries-produce-most-sugar.asp
26. Zakidizaji, H., Bahrami, H., Monjezi, N., & Shiekhdavoodi, M. (2019). Modeling of the variables that influence sugarcane yield using C5. 0 and QUEST decision tree algorithms. *Journal of Agricultural Machinery*, *9*(2), 469-484. https://doi.org/10.22067/jam.v9i2.69712